Impact of the Iteration Length on the Software Quality

Dobromir M. Dinev(✉)

New Bulgarian University, 21 Montivideo Street, Sofia, Bulgaria
dobromir@gmail.com

Abstract. The article disentangles the topic of iteration length of the software development and delivery process; a study is presented in the field of software quality that examines the relationship between iteration length and software quality. The study includes ten teams delivering software in the fintech sector. The text of the article holds a conversation on the potential reasons for the improvement and, with a critical voice, points to further aspects, like software maturity, to be examined. At the same time, a clear message for not shortening too extensively the time to develop and deliver is posed, as this is triggering the situation with the constantly evolving complexity of the software products that are delivered.

Keywords: Iteration length · sprint length · software · quality

1 Introduction

Time is one of the most apparent resources available; it is not like the money to which all other resources can be transferred, and yet the money might not buy you time in some situations. Such general settings, which are related to the time, are making this impactor recognised by the Holistic approach to quality theory very interesting for exploration and research. The current paper introduces a study that aims to investigate whether there is any relationship between the cycle length (the iteration or sprint) and the number of defects and the severity of the defects, e.g., on the quality.

Some research, like [1], presents the monthly delivery cycle as "very short." The study in the current article will examine the effects on quality if the cycle time is further shortened.

2 The Research

The study uses data from ten teams in a multinational company that delivers software for the fintech industry. The company desires to stay anonymous. All the teams deliver to the same client. The change is related to the decision to reduce the iteration cycle time by fifty percent. Initially, the iteration cycle (the sprint) was twenty working days, and after the change, it became ten working days. The general variable expected to indicate any impact is the time needed to resolve the found defects. For any of the teams, no major personnel changes were made for the time of the study; thus, Tuckman's theory [2] is excluded from having any significant impact on the main variable of interest.

T. Zlateva and G. Tuparov (Eds.): CSECS 2024, LNICST 609, pp. 136–145, 2025.
https://doi.org/10.1007/978-3-031-84312-9_9

2.1 The Collected Data for "Iteration Cycles' Time"

The data collected in this study is for ten teams for twelve months duration. The data collected during the study is saved in the ex3_cd data frame Table 1 – Data frame one – time-complexity-quality describes the columns in the data frame. The sprints from S1 to S6 have a length iteration cycle of 20 days; this makes the observation period for the first releases six months. The following sprints from S7 to S18 are with the length of ten days, and the observation period is the same.

Table 1. Data frame one: time-complexity-quality.

Variable	Description
team	The team is an identifier and a factor. Having a factor value from t1, t2,.. t10
severity	Defect's severity, a factor with values: Cosmetic, Low, Normal, Major, and Critical
iter_lenght	A number marking the length of the iteration cycle, 20 or 10
time_spent	The time spend in working hours to resolve the defect
sprint	The particular sprint (cycle iteration). Having a factor value from S1, S2,.. S18
rel_order	The relative order of appearance of the defect for each team
rel_order_two	The relative order two used for the charts

2.2 Statistical Diagnostics and Tests for "Iteration Cycles' Time"

Table 2 – Defects per sprint, per severity – holds the information posed by the table's name. Several findings are immediately evident when looking at the table. First, most of the defects are of normal severity (67.13%). The major and critical defects are almost one-fourth of the total number of reported defects (25,74%, made from 17,16% for the Major severity defects and 8,58% for the critical ones). Second, after the change has been enforced, the number of defects decreases. The decrease seems to be drastic after sprint seven; still, a fact should be considered: the time of the sprint is half, starting from the same sprint seven, so this might be a source of the present change when looking at the raw data. To eliminate this, the sprints starting with sprint seven will be coupled. For instance, sprint seven will be paired with sprint eight, sprint nine will be paired with sprint ten, etc. Table 3 - Defects per sprint, per severity - time corrected; holds the rearranged data. With the time correction in place, it is visible that the defects are decreasing in the second sprint after the change has been made.

Table 2. Defects per sprint, per severity.

Number of defects / Severity	Cosmetic	Low	Normal	Major	Critical	
S1		10	109	31	21	171
S2		12	117	38	19	186
S3		18	120	48	18	204
S4	1	6	109	32	14	162
S5		10	130	45	27	212
S6		9	109	31	24	173
S7	1	11	90	38	12	152
S8		2	47	15	5	69
S9	1	5	43	9	10	68
S10		5	62	8	3	78
S11	1	8	50	6	1	66
S12	3	10	56	9	4	82
S13		5	53	9	2	69
S14	1	4	37	7	3	52
S15		1	53	3	1	58
S16	1	4	56	6	3	70
S17	1	5	53	3	2	64
S18		6	36	2	1	45
	10	131	1330	340	170	1981

There were no considerable changes in the teams regarding the number of people at the time of the study. Furthermore, no other changes were made to the process or technological stack. The load of the teams with new features is kept at a similar pace; a neglectable seasonality is present for the majority of the people. One more reason to consent to the hypothesis that the enforced change is causing the present shift in the number of defects is the seniority of the testers in the agile teams; only seniors are making the test team in the project, and as much as there is always a learning curve, the people passed the same selection procedure for each of the functions (business analysis, development, testing, DevOps). The developers across the teams were of different seniority; few developers were promoted at the end of the study.

After the change has been enforced, it is evident that the number of defects crossing the hundred working hours psychological border is none. (Fig. 1 - Time to resolve defects before and after the change.) Further, visual evidence for the decreasing number of defects is seen in the second chart in Fig. 1, which shows the number of defects across the teams is also decreasing; this is visible by the less frequent points appearing in the interval between forty and sixty in the relative order dimension. Still, one more question remains: Why is there no stronger visible shrink of the number of defects? It is visible the teams are not contributing equally to this shrinkage. The total number of defects logged for the first six months of the study is 1108 defects; after the change has been introduced, the number of defects is 873 for the same length of time. This hit the point that further examination should be put in place to determine the causes of these decrees. Before

Table 3. Defects per sprint, per severity – time corrected.

Severity / Number of defects	Cosmetic	Low	Normal	Major	Critical	
S1		10	109	31	21	171
S2		12	117	38	19	186
S3		18	120	48	18	204
S4	1	6	109	32	14	162
S5		10	130	45	27	212
S6		9	109	31	24	173
S7	1	13	137	53	17	221
S8	1	10	105	17	13	146
S9	4	18	106	15	5	148
S10	1	9	90	16	5	121
S11	1	5	109	9	4	128
S12	1	11	89	5	3	109
	10	131	1330	340	170	1981

committing to this consequent examination, let's see first if this change is statistically significant considering the time spent. In order to select the correct statistical hypothesis testing method, it is required to check the normality of the data that has been collected.

First, this matter will be approached visually. The violin display (Fig. 2 - Violin display of the collected data after and before the change.) of the collected data, after and before the change shows a long tail distribution of the data. According to the Q-Q plots (Fig. 3 – Q-Q plot of time_spend variable, vertically, horizontally the quantiles.), it is detectable that a none-normally distributed of the data is present.

To confirm further the none-normal distribution of the data that has been collected during the study and stored in time_spend variable, next tests for normality are to be performed for the same variable: Shapiro-Wilk ($W = 0.88293$, p-value $< 2.2e^{-16}$), Anderson-Darling($A = 63.454$, p-value $< 2.2e^{-16}$), and Jarque-Bera (X-squared $=$ 942.12, df $= 2$, p-value $< 2.2e^{-16}$). All tests have a p-value below 0,05%; they all reject the hypothesis the data stored in the time_spend variable is with normal distribution. In this case, the Mann-Whitney-Wilcoxon test will be applied. Before that, check the summary statistics will be performed (Table 4 - Summary statistics before Mann-Whitney-Wilcoxon):

Fig. 1. Time to resolve defects before and after the change.

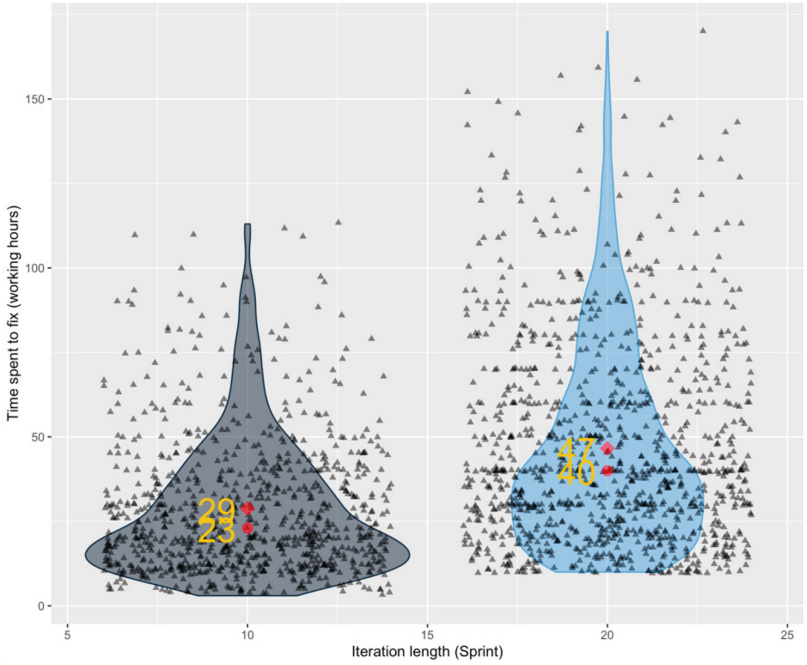

Fig. 2. Violin display of the collected data after and before the change.

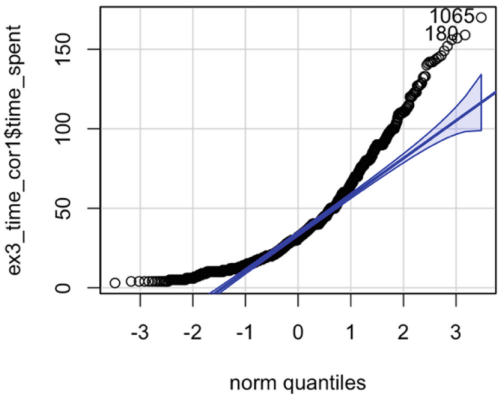

Fig. 3. Q-Q plot of time_spend variable, vertically; horizontally the quantiles.

Table 4. Summary statistics before Mann-Whitney-Wilcoxon.

	iter_lenght	n	median	irq
1	10	873	23	25
2	20	1108	40	36

Then, calculating the test: (Table 5)

Table 5. Mann-Whitney-Wilcoxon test results.

variable	group1	group2	n1	n2	statistic	p
time_spent	10	20	873	1108	290668	$1.2e^{-52}$

Last checking the effects size:

Table 6. Mann-Whitney-Wilcoxon test, effects.

	group1	group2	effsize	n1	n2	magnitude
time_spent	10	20	0.343	873	1108	moderate

Based on the run diagnostics and statistical tests, it might be concluded that the medina for the time spent, in case the iteration length was twenty working days is 40 (IRQ = 36), whereas the median for time spent for the iteration length of ten working days is 23 (IRQ = 25). The Wilcoxon test showed that the difference was significant ($p = 1{,}2e^{-52}$, effect size = 0,343).

Table 7. Defects per sprint, per severity – before the change.

Severity		Sprints						R. total %
		S1	S2	S3	S4	S5	S6	
Cosmetic	Count	0	0	0	1	0	0	1
	Chi-sq	0,154	0,168	0,184	4,986	0,191	0,156	-
	Row %	0,000	0,000	0,000	100,000	0,000	0,000	0,090
	Col. %	0,000	0,000	0,000	0,617	0,000	0,000	-
	Total %	0,000	0,000	0,000	0,090	0,000	0,000	-
Low	Count	10	12	18	6	10	9	65
	Chi-sq	0,000	0,109	3,041	1,292	0,477	0,130	-
	Row %	15,385	18,462	27,692	9,231	15,385	13,846	5,886
	Col. %	5.848	6,452	8,824	3,704	4,714	5,202	-
	Total %	0,903	1.083	1,625	0,542	0,903	0,812	-
Noramal	Count	109	117	120	109	130	109	694
	Chi-sq	0,033	0,002	0,473	0,559	0,058	0,004	-
	Row %	15,706	16,859	17,291	15,706	18,732	15,706	62,635
	Col. %	63,743	62,903	58,824	67,284	61,321	63,006	-
	Total %	9,838	10,560	10,830	9,838	11,733	9,838	-
Major	Count	31	38	48	32	45	31	225
	Chi-sq	0,400	0,001	1,043	0,024	0,088	0,486	-
	Row %	13,778	16.889	21,333	14,222	20,000	13,778	20,307
	Col. %	18.129	20,430	23,529	19,753	21,226	17,919	-
	Total %	2,798	3,430	4,332	2,888	4,061	2,798	-
Critical	Count	21	19	18	14	27	24	123
	Chi-sq	0,214	0,132	0,953	0,882	0,510	1,197	-
	Row %	17,073	15,447	14,634	11,382	21,951	19,512	11,101
	Col. %	12.281	10,215	8,824	8,642	12,736	13,873	-
	Total %	1,895	1,715	1,625	1,264	2,437	2,166	-
Column	Total	171	186	204	162	212	173	1108
	%	15,433	16,787	18,412	14.621	19,134	15,614	

Let's now examine the severity of the defects in the sprints, i.e., in time. This will help synthesise the answer to the question: Is there a change in the quality if the delivery time is shrunken? Further, if there is an impact, is it positive or negative? Table 7 holds the information for the first six months and Table 8 for the following six months after the change.

Table 8. Defects per sprint, per severity – after the change.

Severity		Sprints						R. total %
		S7	S8	S9	S10	S11	S12	
Cosmetic	Count	1	1	4	1	1	1	9
	Chi-sq	0,717	0,170	4.01	20,049	0,077	0,014	-
	Row %	11,111	11,111	44,444	11,111	11,111	11,111	1,031
	Col. %	0,452	0685	2,703	0,826	0,781	0,917	-
	Total %	0,115	0,115	0,458	0,115	0,115	0,115	-
Low	Count	13	10	18	9	5	11	66
	Chi-sq	0,823	0,098	4,146	0,002	2,260	0,924	-
	Row %	19,697	15,152	27,273	13,636	7,576	16,667	7,560
	Col. %	5,882	6,849	12,162	7,438	3,906	10,092	-
	Total %	1,489	1,145	2,062	1,031	0,573	1,260	-
Noramal	Count	137	105	106	90	109	89	636
	Chi-sq	3,579	0,017	0,031	0,039	2,660	1,158	-
	Row %	21,541	16,509	16,667	14,151	17,138	13,994	72,852
	Col. %	61,991	71,918	71,622	74,380	85,156	81,651	-
	Total %	15,693	12,027	12,142	10,309	12,486	10,195	-
Major	Count	53	17	15	16	9	5	115
	Chi-sq	19,601	0,259	1,037	0,000	3,665	6,100	-
	Row %	46,087	14,783	13,043	13,913	7,826	4,348	13,173
	Col. %	23,982	11,644	10,135	13,223	7,031	4,587	-
	Total %	6,071	1,947	1,718	1,833	1,031	0,573	-
Critical	Count	17	13	5	5	4	3	47
	Chi-sq	2,188	3,361	1,106	0,352	1,213	1,402	-
	Row %	36,170	27,660	10,638	10,638	8,511	6,383	5,384
	Col. %	7,692	8,904	3,378	4,132	3,125	2,752	-
	Total %	1,947	1,489	0,573	0,573	0,458	0,334	-
Column	Total	221	146	148	121	128	109	873
	%	25,315	16,724	16,953	13,860	14,661	12,486	

Before the change (Table 9), i.e., for the first six sprints, the total percentage of cosmetic was 0,09%; after the change, this figure slightly changed to 1,31%. For the low severity defects, the figures show a similar small change from 5,86% to 7,76%. A shift of nearly 10% has been introduced for the normal severity defects; this figure here shifted from 62,63% to 72,85%. The major change was from 20,30% to 13,17%, and the defects set to be critical by the testers changed from 11,10% to 5,38%. By the

Table 9. Summary table of before and after the change by severity.

Severity	Before, %	After, %
Cosmetic	0,09	1,31
Low	5,86	7,76
Normal	62,63	72,85
Major	20,30	13,17
Critical	11,10	5,38

figures in Table 6, it might be concluded that there is a decrease in the major and critical defects in the second six months of the study, the change of 13,34% from the major and critical severity defects are mainly moved to normal severity defects (10,22%), 1,22% are transferred to the cosmetic severity defects for the time after the change, and 1,9% respectively to the low severity. Consequently, this shows the change in the quality of the code of the software solution and, in general, shows a situation with fewer defects delivered to production; the delivered quality is higher.

The percentage changes are not enough to make a scientific conclusion; this is why statistical proof for the change is required. This is a proportion change, and to examine it, the proportion test will be applied. After the run of the proportional test for each of the present defect severities, the results in Table 10 were acquired.

Table 10. Results from two-sided proportion tests by severity.

Severity	R code	p-value
Cosmetic	prop.test(x = c(1, 9), n = c(1108, 873))	0,008956
Low	prop.test(x = c(65, 66), n = c(1108, 873))	0,1571
Normal	prop.test(x = c(694, 636), n = c(1108, 873))	1,953e-06
Major	prop.test(x = c(225, 115), n = c(1108, 873))	3,777e-05
Critical	prop.test(x = c(123, 47), n = c(1108, 873))	9,433e-06

Table 10 - Results from two-sided proportion tests by severity; holds the results from the run two-sided proportional test with H_0 that the proportions are equal. The p-values for normal, major, and critical severity are showing that the proportions are different and, in those cases, the H_0 is rejected, i.e., *there is an improvement to the quality considering the severity of the reported defects.* The test for low severity is insignificant as the p-value here is 0,1571.

3 Conclusion

The conducted study showed that the iteration's length impacts the quality of the delivered software solution. The number of defects has decreased in the time after the iteration's length has been cut in half. The total number of defects before the change was introduced has been decreasing from one thousand, one hundred and eight to eight hundred seventy-three. A statistically significant change in the proportions has been identified in favour of decreasing the severity of the defects; at the same time, the time spent for fixing is shrunken drastically, which shows a business impact.

A further study may be conducted in relation to iteration's (sprint) length; for instance, if further shrinking the time for delivery is present, will it be the same, as proven by the current study, the tendency of improvement in quality be sustained. The hypothesis of the author of the current work is that this will not happen, as the time for the functions will be shortened, this will request the functionality delivered to be fragmented more; unfortunately, this fragmentation has its limitations, especially in the context of the growing complexity of the technical solutions. A study that considers the complexity of the process followed and /or the code is suggested.

A more detailed study should be conducted on the accumulation of defects when the length of the iteration is increased.

The study in the current article proves one of the central statements related to lean: "Of all the ideas of Lean, batch size reduction - is the most important economically" [3]. Data for this is not shared; no doubt increasing the quality and making fast delivery to production is having a positive financial impact.

References

1. Van Oorschot, K.E., Van Wassenhove, L.: Under pressure: the effects of iteration lengths on agile software development performance. Project Management J. **49**(3), 875697281880271 (2018)
2. Tuckman, B.W: Developmental sequence in small groups. Psychological Bulletin. **63**(6), 384–399. https://doi.org/10.1037/h0022100. PMID 14314073 (1965)
3. Reinertsen, D.G.: The Principles of Product Development Flow: Second Generation Lean Product Development. Celeritas Publishing (2009)